# LLM ADVERTISEMENT BASED ON NEURON AUCTIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

As Large Language Models (LLMs) transition into conversational agents, generative advertising emerges as a crucial monetization strategy. However, embedding advertisements within unstructured LLM outputs introduces a critical trilemma: balancing advertiser payoffs, platform revenue, and user experience. Existing methods, such as prompt injection or rigid position slots, disrupt semantic coherence and lack a parametric framework for independent control, rendering rigorous mechanism design intractable. To bridge this gap, we introduce *Neuron Auctions*, a novel paradigm that shifts the auction object from the surface text space to the LLM's internal representations. Leveraging mechanistic interpretability, we identify brand-specific feed-forward network (FFN) neuron sets with low overlap in the evaluated settings. We use neuron counts as fine-grained intervention budgets that serve as auctionable commodities. Building on this computational carrier, we design a menu-based auction mechanism and establish an approximate strategy-proofness guarantee under bounded cross-advertiser interference. The mechanism is further optimized for platform revenue while explicitly accounting for user utility, thereby discouraging overly aggressive interventions. Direct empirical comparisons against simple prompt-level and fixed-position ad insertion baselines show more responsive and consistent bid-dependent control over predicted click intent, while menu-training experiments demonstrate a controllable trade-off among platform revenue, advertiser utility, and user utility.

## 1 INTRODUCTION

The rapid adoption of Large Language Models (LLMs) is driving a fundamental paradigm shift in information retrieval, transitioning traditional search engines into conversational AI agents. This evolution naturally necessitates a corresponding shift in commercial monetization towards generative advertising (Duetting et al., 2024; Zhao et al., 2025). Advertising in LLMs is embedded in an autoregressive generation process that jointly determines what information is presented and how it is surfaced. This introduces a new mechanism design challenge involving three stakeholders: advertisers, users, and the platform. Advertisers seek to maximize their Click-Through Rates (CTR) and Return on Investment (ROI) through highly visible recommendations; users value helpful, coherent, and high-quality responses; and platforms aim to maximize revenue while preserving user retention and satisfaction. A central question is reconciling these commercial and experiential objectives within the unstructured, autoregressive generation process of LLMs, mitigating the risk of models forsaking user welfare to heavily promote sponsored options (Wu et al., 2026).

Existing approaches provide important first steps, but they do not fully address this challenge. Extensions of classical search advertising mechanisms, such as Generalized Second-Price (GSP) auctions (Edelman et al., 2007; Varian, 2007), are built on position-based allocation and naturally assume predefined advertising slots. While effective in traditional search settings, such slot-based designs are less suited to the fluid and context-dependent nature of LLM responses, where advertisements must be integrated natively into generated content. More recent work has explored advertising mechanisms tailored to LLM-generated content, including position-, segment-, and prompt-based approaches (Balseiro et al., 2026; Hajiaghayi et al., 2024), offering a more flexible way to influence model outputs. However, these approaches typically do not provide a parametric mechanism for assigning fine-grained, independent recommendation strengths to competing brands, nor can they explicitly control user utility degradation. In addition, existing methods often inherit standard pricing rules such as GSP or VCG, leaving open the question of how to optimize platform revenue in the generative advertising setting.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We trace these limitations to a central difficulty in generative advertising: the choice of the auctionable object. In classical search advertising, the auctionable object is a position; in prompt-based approaches, it is an instruction or insertion opportunity. Both are defined in the surface text space, where control over generation is often coarse, entangled with response quality, and difficult to parameterize. In contrast, effective monetization in LLMs requires a computational carrier that is advertiser-specific, independently controllable, compatible with the model's native generation process, and, crucially, equipped with a measurable intervention budget that can be translated into allocation and pricing.

To this end, we shift the auction paradigm from surface-level text to the model's internal representation space. We introduce ***Neuron Auctions***, in which the auctionable commodities are selected feed-forward network (FFN) neurons. Building on recent advances in mechanistic interpretability (Dai et al., 2022; Turner et al., 2023), we identify subsets of "brand-specific neurons" whose activations modulate the model's propensity to recommend particular advertisers. This representation-space formulation enables fine-grained control over recommendation strength through explicit intervention parameters, including the number of activated top-$k$ neurons and their amplification factor $\lambda$, while allowing the final response to be generated naturally by the LLM.

A key property enabling this design is separability. Our selected neuron sets exhibit little overlap, and we further assess cross-advertiser effects on predicted click intent by comparing single-advertiser and joint interventions. These structural and behavioral analyses motivate our bounded-interference formulation. The resulting approximate separability makes FFN activations a natural computational carrier for generative advertising: *advertiser-specific interventions can be parameterized, allocated, and priced without forcing advertisements into rigid slots or disrupting the fluidity of generation.*

This structure further provides the foundation for a mechanism design framework. The intervention budget—specified by the number of intervened neurons—acts as a controllable proxy for recommendation intensity. Empirically, increasing the intervention budget generally strengthens the target brand's presence in generated responses, supporting the use of neuron counts as a controllable measure of intervention intensity. This parametric relationship allows us to treat neurons as auctionable commodities and to design a *menu-based auction mechanism*. In our framework, the platform offers each advertiser a menu of prices and expected intervention levels, allowing advertisers to select the degree of recommendation strength that best matches their value. Each advertiser's menu is independent of its own reported value, yielding approximate strategy-proofness under bounded cross-advertiser interference, while the menu prices are optimized through a neural network with respect to the platform objective.

We incorporate both platform revenue and a user utility into this objective, so that overly aggressive interventions are priced or discouraged when they degrade response quality. Our experiments show that our method can achieve higher platform revenue than the VCG baseline, and differs from approaches that approximate VCG-style payments through computationally intensive multi-fidelity optimization (Clarke, 1971; Groves, 1973; Vickrey, 1961; Liu et al., 2026). Together, the independence and priceability of neuron-level interventions allow *Neuron Auctions* to address the generative advertising trilemma: improving advertiser visibility, preserving users' conversational experience, and optimizing platform revenue within a unified framework.

## 2 Related Work

**Mechanism Design for Generative Advertising** Traditional search advertising has been heavily influenced by the Generalized Second-Price (GSP) auction (Edelman et al., 2007; Varian, 2007) and the truthful Vickrey-Clarke-Groves (VCG) mechanism (Clarke, 1971; Groves, 1973; Vickrey, 1961). However, integrating these classic mechanisms into generative AI requires a paradigm shift: traditional position-based models are incompatible with the unstructured, autoregressive nature of LLMs, and calculating exact VCG payments demands computationally prohibitive counterfactual evaluations. To directly address this computational bottleneck, Liu et al. (2026) coupled VCG incentives with multi-fidelity optimization to maximize expected social welfare under strict generative budgets while establishing formal guarantees for approximate strategy-proofness. Parallel to such computational optimizations, recent literature has actively redefined the auction object itself. Duetting et al. (2024) introduced the foundational token auction model via single-dimensional bids, where advertisers bid token-by-token to insert their preferred words. Subsequently, ad auctions have been

customized for Retrieval-Augmented Generation (RAG) systems (Hajiaghayi et al., 2024), and position auctions in continuous AI-generated content have been formalized by explicitly modeling substitution effects through multinomial logit and cascade user behaviors (Balseiro et al., 2026). To further mitigate latency and privacy issues, other frameworks either treat the LLM output distribution as the auction object via reward-preference optimization (Zhao et al., 2025), or completely decouple ad insertion from generation by utilizing semantic "genres" as a proxy for VCG auctions (Xu et al., 2026).

Yet, introducing competing commercial objectives poses severe alignment risks, as models may forsake user welfare to heavily promote sponsored options (Wu et al., 2026). More importantly, whether utilizing decoupled frameworks or token-level bidding, relying on surface-level text insertions or external prompt manipulations fundamentally disrupts the natural discourse and semantic coherence of the generated response. Unlike these methods, our approach shifts the auction paradigm directly to the internal representation space of the LLM, maintaining approximate strategy-proofness while explicitly incorporating user utility degradation through native neuron interventions.

**Influencing LLMs and Internal Intervention** To achieve fine-grained, independent control over brand recommendations without prompt manipulation, we must consider how to best influence modern LLMs. During post-training, parameter-efficient fine-tuning (PEFT, e.g., LoRA) inserts small modules that achieve strong effects (Hu et al., 2022; Pfeiffer et al., 2021). However, deliberate curation of duplicated data, full pretraining, RLHF, or DPO hard-wire behaviors (Ouyang et al., 2022; Rafailov et al., 2023), which are impractical for rapid, per-query, multi-stakeholder advertising. At inference, system prompts can be brittle, and plug-and-play decoding steers generation via auxiliary scores at a noticeable latency cost (Dathathri et al., 2020; Krause et al., 2021).

Therefore, our intervention mechanism leverages recent breakthroughs in mechanistic interpretability. Previous work has demonstrated that feed-forward network (FFN) layers function as key-value memories (Dai et al., 2022). Recent techniques like Activation Addition (ActAdd) and Sparse Autoencoders (SAEs) edit interpretable features on-the-fly, offering continuous control knobs on influence strength without weight updates (Turner et al., 2023; O'Brien et al., 2024; Koriagin et al., 2025). Our method is highly synergistic with contrastive activation addition (Panickssery et al., 2023), utilizing cloze-prompt differences to isolate brand neurons. By extending these steering techniques into an economic setting, we use advertiser-specific neuron intervention budgets as auctionable commodities.

## 3 Neuron Auction Framework

We consider an LLM platform that mediates $n$ advertisers, where advertiser $a$ is associated with a brand $b_a$. Our goal is to define an auctionable object that can modulate the model's tendency to recommend a sponsored brand while preserving the natural autoregressive generation process. To this end, rather than allocating surface-level positions or appending advertising prompts, we intervene on a small set of brand-related feed-forward network (FFN) neurons in the LLM.

For each brand $b$, we first identify a ranked set of brand-specific neurons using gradient-based attribution. Let $\Omega_b(k)$ denote the top-$k$ neurons most strongly associated with brand $b$. An intervention option for advertiser $a$ is parameterized by the number of intervened neurons, $k_a$. During response generation, the input prompt remains unchanged. Instead, for each neuron $i$ in layer $l$, we modify its FFN activation as

$$\tilde{h}_i^{(l)} = \begin{cases} \lambda h_i^{(l)}, & i \in \Omega_{b_a}(k_a), \\ h_i^{(l)}, & i \notin \Omega_{b_a}(k_a). \end{cases}$$

Here, $h_i^{(l)}$ is the original intermediate activation, $\tilde{h}_i^{(l)}$ is the intervened activation, and $\lambda$ is an amplification coefficient. Thus, the allocation variable $k_a$ directly controls the intervention degree.

This representation-space formulation has three properties that make it suitable for auction design. First, the intervention is advertiser-specific: each bidder is associated with a brand-dependent ranked neuron set. As we show later, our construction of brand-specific neuron sets is also approximately separable across advertisers, enabling largely independent control of competing brand interventions. Second, the allocation is measurable and ordered: larger values of $k_a$ correspond to stronger intervention levels, allowing the platform to construct menus over recommendation intensity. Third, because

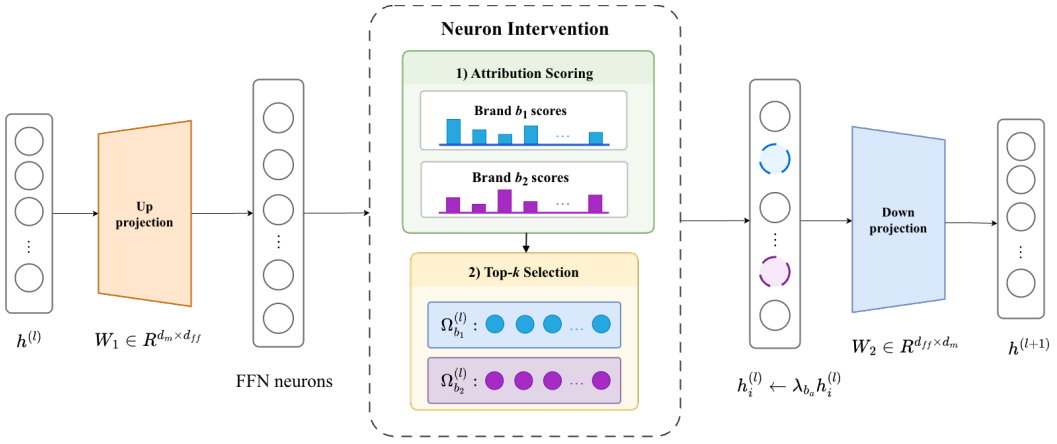



Figure 1: Overview of the proposed neuron intervention pipeline in the FFN at layer $l$. $\Omega^{(l)}_{b_a}$ stands for the Top-$k$ Neurons chosen to be amplified for bidder $b_a$ in Layer $l$, and $\lambda_{b_a}$ is the intervention coefficient.

the response is still generated autoregressively by the LLM, the intervention avoids rigid ad slots or direct prompt manipulation and better preserves the fluency and coherence of the generated answer.

### 3.1 Intervention Method

**Offline attribution scoring.** The first step is to identify neurons that are selectively associated with each brand. For every brand $b$, we calculate an attribution score for each FFN neuron, where a larger score indicates that the neuron more strongly promotes brand $b$. The resulting ranked neuron list defines the candidate intervention set from which the platform later constructs auction options. The key challenge is specificity: a useful neuron should be highly responsive to the target brand while being weakly responsive to competing brands.

We adopt a cloze-style attribution strategy to calculate brand-level scores. For each brand $b$, we construct $M$ semantically similar but lexically diverse cloze prompts, denoted by $\mathcal{P}_b = \{p^{(b)}_1, \ldots, p^{(b)}_M\}$. All prompts are designed to elicit the same brand concept. Using a single prompt can introduce substantial contextual noise, making attribution scores sensitive to superficial prompt variations. Averaging over multiple synonymous cloze prompts reduces this variance and emphasizes features consistently associated with the brand itself. Examples of cloze prompts are provided in Appendix B.

However, neuron intervention methods (Dai et al., 2022) typically define attribution with respect to a single target token. This is insufficient for brand names that are tokenized into multiple sub-tokens. For example, Qwen3 tokenizes "Hyatt" as `['Hy', 'att']`. We therefore use the log-probability of the full brand token sequence as the attribution target.

Let $y^{(b)} = (y_1, \ldots, y_T)$ be the tokenized brand and $p'_t = p \,\|\, y_{<t}$ be the prefix for predicting token $y_t$. For cloze prompt $p$, the attribution score of neuron $(l, i)$ is defined as

$$A^{(l)}_i(b; p) = \sum_{t=1}^{T} h^{(l)}_i(p'_t) \int_0^1 \frac{\partial \log P\left(y_t \mid p'_t; \alpha h^{(l)}_i(p'_t)\right)}{\partial \left(\alpha h^{(l)}_i(p'_t)\right)} \, d\alpha. \tag{1}$$

Here, $h^{(l)}_i(p'_t)$ is the FFN intermediate activation at the last token position of $p'_t$. The log-probability form decomposes the attribution of a multi-token brand into additive token-wise terms. It also avoids the numerical instability caused by multiplying several token probabilities.

In addition, to further identify neurons that are truly brand-specific, we construct for each brand $b$ a set of competing brands, denoted by $\mathcal{N}(b) = \{b^-_1, \ldots, b^-_K\}$, and adopt a contrastive strategy. Specifically, we first average the attribution of the target brand across multiple cloze prompts:

$$S_i^{(l)}(b) = \frac{1}{|\mathcal{P}_b|} \sum_{p \in \mathcal{P}_b} A_i^{(l)}(b; p),$$

and then subtract the average attribution of the same neuron to competing brands:

$$\tilde{S}_i^{(l)}(b) = S_i^{(l)}(b) - \frac{1}{|\mathcal{N}(b)|} \sum_{b^- \in \mathcal{N}(b)} S_i^{(l)}(b^-).$$

The larger this score is, the more likely neuron $(l, i)$ is to be a "true positive neuron" for the target brand, meaning that it tends to promote the generation of the target brand rather than generally supporting the outputs of brands in the same category.

Since the exact computation of the integral is expensive, we approximate it using a Riemann sum with $m$ discrete steps. The attribution score used in practice is therefore written as

$$A_i^{(l)}(b; p) \approx \sum_{t=1}^{T} \frac{h_i^{(l)}(p'_t)}{m} \sum_{r=1}^{m} \left. \frac{\partial \log P\Big(y_t \mid p'_t; \alpha h_i^{(l)}(p'_t)\Big)}{\partial \Big(\alpha h_i^{(l)}(p'_t)\Big)} \right|_{\alpha = r/m}.$$

In our experiments we set $m = 20$, achieving a good balance between computational cost and attribution stability. Through the above procedure, for any brand, we are able to reliably identify its corresponding "true positive neurons".

**Online Neuron Selection and Intervention**. After obtaining neuron attribution scores for each brand, we further perform a second-stage filtering procedure based on *brand specificity*. Suppose that there are $N$ bidders participating in one auction round, denoted by $\{b_1, \ldots, b_N\}$. For a candidate neuron $i$ associated with brand $b_j$, we define its online ranking score as

$$\mathrm{Attr}_i^{(l)}(b_j) = \tilde{\mathrm{S}}_i^{(l)}(b_j) - \frac{\sum_{b' \neq b_j} \left| \tilde{\mathrm{S}}_i^{(l)}(b') \right|}{N - 1}.$$

Specifically, by subtracting the absolute attribution scores of all non-target brands, we penalize neurons that are also highly associated with competing brands, regardless of whether their effects are promotional or inhibitory. Based on this criterion, we rank neurons separately for each brand and select the top-$k$ neurons to form the final intervention set. During generation, we only amplify the selected neurons at the last token of the current decoding step, which is consistent with the autoregressive decoding process of Transformer decoders. Let $\Omega_b$ denote the target neuron set associated with brand $b$, and let $\lambda_b > 1$ denote the amplification factor. We then intervene on the FFN intermediate activations of the last token as follows:

$$h_i^{(l)} \leftarrow \lambda_b h_i^{(l)}, \quad (l, i) \in \Omega_b.$$

However, when multiple bidders simultaneously inject their brand-specific neurons, independently amplifying all selected neurons often causes the hidden-state norm to grow rapidly, which can further lead to response collapse. To address this issue, we design a *unified-hook* mechanism: we first apply amplification to the target neurons of all brands, and then perform a unified norm rescaling on the modified last-token representation so that its overall $L_2$ norm is restored to the pre-intervention level.

$$h_{new}^{(l)} = \frac{||h_{old}^{(l)}||_2}{||h_{intervened}^{(l)}||_2} \cdot h_{intervened}^{(l)},$$

where $h_{\mathrm{old}}^{(l)}$ denotes the original FFN intermediate activation vector at the last token position in layer $l$, $h_{\mathrm{intervened}}^{(l)}$ denotes the vector after applying all selected neuron amplifications, and $h_{\mathrm{new}}^{(l)}$ denotes the final rescaled vector fed back into the model. This rescaling controls the overall magnitude of the FFN intermediate representation after intervention. Without this normalization, simultaneously amplifying neurons for multiple brands can substantially increase the activation norm, pushing the representation away from the typical activation distribution and causing unstable or repetitive decoding.

In this way, we increase the probability that the target brand is mentioned and recommended in natural responses, while preserving generation stability.

### 3.2 NEURON AUCTIONS

We now define the auction mechanism built on top of neuron-level interventions. The auctionable object is the intervention degree $k$: the number of brand-specific neurons whose activations are amplified during generation. As shown in Section 4.1, larger intervention budgets are generally associated with stronger target-brand presence, making $k$ a quantifiable control parameter for advertising intensity. Building on this intervention parameter, we adopt a menu-based auction representation (Wang et al., 2024b; 2025) allowing the platform to price different levels of recommendation strength directly in the model's representation space.

For each advertiser $i$, the platform offers a menu

$$\mathcal{M}_i(v_{-i};\theta) = \left\{ \left( k_i^{(j)}, p_i^{(j)}(v_{-i};\theta) \right) \right\}_{j=1}^{J}, \tag{2}$$

where each option consists of an allocation level $k_i^{(j)}$ and a corresponding price $p_i^{(j)}$. The allocation $k_i^{(j)}$ specifies how many brand-specific neurons of advertiser $i$ will be amplified, thereby determining the intervention degree and the induced advertising effect. Let $C_i(k_i, k_{-i})$ denote advertiser $i$'s click-through rate (CTR) under the joint intervention profile $(k_i, k_{-i})$. We evaluate menu options using a reference CTR $\bar{C}_i(k_i)$, which is fixed with respect to advertiser $i$'s own reported value. In our implementation, we set $\bar{C}_i(k_i) = C_i(k_i, \mathbf{0})$, keeping the neuron rankings fixed and disabling all other advertisers' interventions. Given the reported values, the platform selects each advertiser's menu option by maximizing its reference utility:

$$u_i^{(j)}(v_{-i}) = v_i \bar{C}_i\big(k_i^{(j)}\big) - p_i^{(j)}(v_{-i};\theta), \qquad j_i^* = \arg\max_j u_i^{(j)}(v_{-i}). \tag{3}$$

The platform then allocates intervention degree $k_i^{(j_i^*)}$ and charges price $p_i^{(j_i^*)}(v_{-i};\theta)$. The allocation levels $\{k_i^{(j)}\}_{j=1}^{J}$ are fixed menu options, while the prices are generated by a neural network parameterized by $\theta$. For each bidder, we use a separate neural network that takes the bids submitted by all other advertisers, $v_{-i}$, as input. We also include an option with zero allocation and a fixed price of zero. This ensures nonnegative reference utility under truthful reporting and $\epsilon$-approximate individual rationality under the bounded-interference assumption below.

*Strategy-Proofness Analysis* We next show that bounded cross-advertiser interference leads to a corresponding approximate strategy-proofness guarantee, with the proof deferred to Appendix I.

**Theorem 3.1** (Approximate Strategy-Proofness)**.** *Suppose that each bidder's value satisfies $v_i \in [0,1]$, each bidder's menu and reference CTR function are independent of its own reported value, and menu options are selected according to the reference-utility maximization rule above. Assume that*

$$\left| C_i(k_i, k_{-i}) - \bar{C}_i(k_i) \right| \le \epsilon$$

*for every bidder $i$ and every feasible intervention profile $(k_i, k_{-i})$. Then, for any misreport $b_i$,*

$$u_i(v_i, v_{-i}; v_i) \ge u_i(b_i, v_{-i}; v_i) - 2\epsilon.$$

*Hence, truthful reporting has regret at most $2\epsilon$, and the mechanism is $2\epsilon$-approximately strategy-proof.*

We examine neuron-set overlap in Appendix F and directly evaluate cross-advertiser effects on predicted click intent in Appendix G. These analyses complement the joint-intervention experiments in Section 4.1 and provide structural and behavioral evidence relevant to the bounded-interference formulation.

*User utility.* Let $q(\{k_i\}_{i=1}^n)$ denote user-side response quality under joint intervention. We use the LLM-based scoring procedure in Appendix A to estimate the CTR proxy and user utility. This term captures whether the generated answer remains helpful, coherent, and responsive to the user's query after neuron interventions. Both the click-through function $C(\cdot)$ and the user-quality function $q(\cdot)$ are estimated using the LLM-based scoring procedure described in Appendix A.1, which combines dimension-specific evaluation prompts, evidence-before-score reasoning, few-shot human-calibrated exemplars, and deterministic decoding to provide stable and interpretable proxy measurements.

The platform optimizes prices to balance commercial revenue and user experience. Since advertisers jointly affect response quality, we optimize menu prices to balance revenue and quality:

$$\mathcal{J}(\theta) = \sum_i \sum_j z_i^{(j)} p_i^{(j)}(v_{-i};\theta) + w_{\text{user}} \mathbb{E}_{j_i \sim z_i} \left[ q\left( \{k_i^{(j_i)}\}_{i=1}^n \right) \right]. \tag{4}$$

Here, $z_i^{(j)} = \mathrm{SoftMax}_j(u_i^{(1)}/\tau, \ldots, u_i^{(J)}/\tau)$ provides a differentiable relaxation of the discrete menu choice, where $\tau = 0.03$; the original $\arg\max$ rule is used at inference. The weight $w_{\text{user}}$ controls the trade-off between platform revenue and user utility.

# 4 EXPERIMENTS

We design experiments to answer the following questions: (1) Can neuron-level interventions provide controllable and advertiser-specific advertising effects in multi-advertiser settings? (Sec. 4.1) (2) Can the learned menu mechanism effectively regulate the trade-off among platform revenue, advertiser utility, and user utility? (Sec. 4.2) (3) How does Neuron Auctions compare with prompt-based and slot-based advertising methods? (Sec. 4.3)

## 4.1 INTERVENTION EXPERIMENTS

We conduct intervention experiments on **100 two-bidder combinations** across three large language models, Qwen3-4B (Yang et al., 2025), DeepSeek-R1-8B (Guo et al., 2025), and Llama-3-8B (Grattafiori et al., 2024), as well as on **20 three-bidder combinations** on Qwen3-4B. In all experiments, both attribution and intervention are performed in the FFN space.

For each two-bidder combination $(b_1, b_2)$, we manually design $p = 3$ prompts whose semantic contexts are naturally compatible with both brands. We first compute the contrastive attribution scores for the two advertisers separately and rank neurons accordingly, then select the top-$k$ neurons for each advertiser and perform joint intervention during generation using the unified-hook mechanism. In the two-bidder setting, we sweep the intervention budgets over $k_1, k_2 \in \{0, 100, 200, \ldots, 800\}$. For each combination of three-bidder $(b_1, b_2, b_3)$ in Qwen3-4B, we follow the same procedure. Specifically, we design prompts that are semantically compatible with all three advertisers, compute attribution scores for each advertiser independently, and jointly intervene on their selected neurons during generation. To control the overall generation cost, in the three-bidder setting we use a slightly smaller intervention range: $k_1, k_2, k_3 \in \{0, 100, 200, \ldots, 600\}$.

In Appendix D, we evaluate 10 two-bidder pairs on Qwen3-4B using the Q4 click-intent score defined in Appendix A. The results show that larger own-brand intervention budgets are generally associated with higher Q4 scores, qualitatively consistent with the hit-count results. This consistency motivates using hit count as a lightweight measure of brand presence in the larger-scale intervention experiments below, reducing the cost of LLM-based evaluation. We retain the Q4-based CTR proxy for menu evaluation.

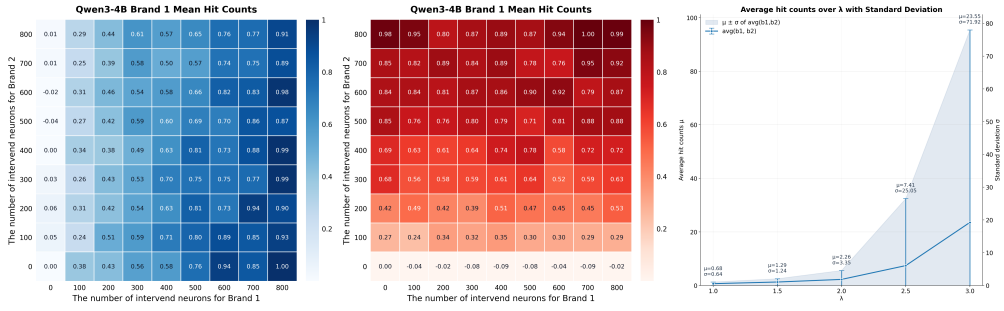
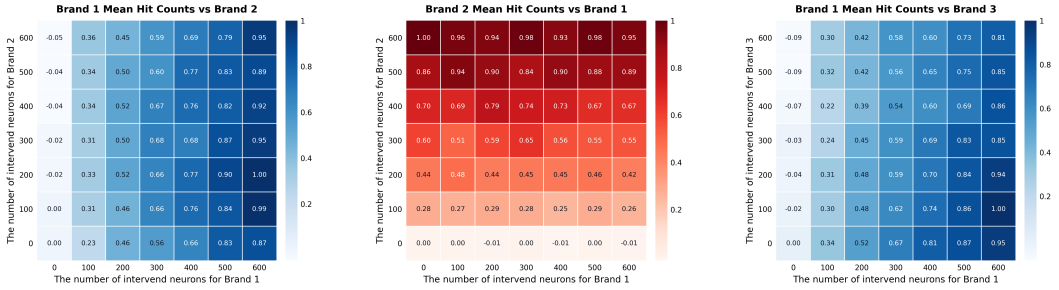
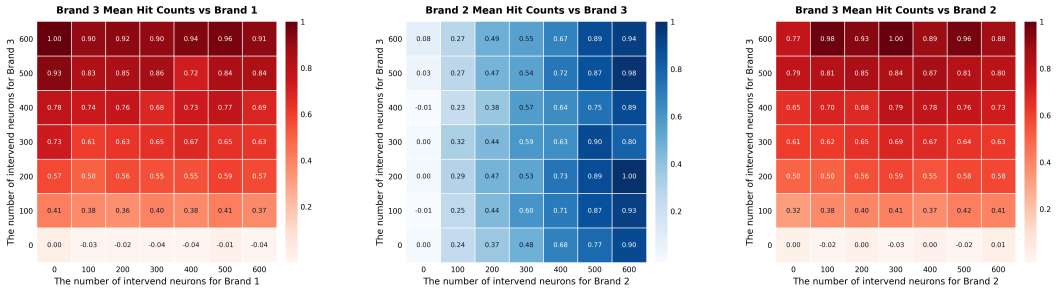



Figure 2: Two-bidder intervention results on Qwen3-4B. The left and middle panels show normalized hit-count changes for the two bidders over intervention budgets $(k_1, k_2)$, while the right panel reports the ablation over $\lambda$ with mean and standard deviation. Heatmaps are averaged across bidder pairs and normalized relative to the $(0, 0)$ baseline.

The results of the two-bidder and three-bidder experiments on Qwen3-4B are summarized in Figure 2 and Figure 3. In both settings, we observe that increasing the intervention budget of a target advertiser consistently strengthens its presence in the generated responses, while cross-advertiser interactions remain secondary. For the two-bidder setting, the corresponding heatmaps for DeepSeek-R1-8B and Llama-3-8B also exhibit similar results, and their heatmaps are provided in Appendix C.




(a) Brand 1 hit counts of $(k_1, k_2)$. (b) Brand 2 hit counts of $(k_1, k_2)$. (c) Brand 1 hit counts of $(k_1, k_3)$.




(d) Brand 3 hit counts of $(k_1, k_3)$. (e) Brand 2 hit count of $(k_2, k_3)$. (f) Brand 3 hit count of $(k_2, k_3)$.

Figure 3: Heatmaps of brand hit counts under joint neuron intervention for three bidders on Qwen3-4B. Each subfigure shows the normalized hit counts of one bidder on a two-bidder intervention plane.

As for the intervention coefficient $\lambda$, we conduct an ablation study on 30 two-bidder combinations using Qwen3-4B, as shown in Figure 2. We find that when $\lambda \leq 1.5$, the intervention effect is relatively weak and the target brands are not sufficiently promoted in the generated responses. In contrast, when $\lambda = 3.0$, the model frequently collapses into repetitive outputs of the target brand names, which leads to sharp increases in both the average hit count and its variance. Therefore, balancing intervention effectiveness and generation stability, we use a shared intervention coefficient $\lambda = 2.0$ for all bidders in the subsequent experiments.

### 4.2 Menu Training Experiments

We next evaluate the learned menu mechanism through representative menu options and the trade-off across different $w_{\text{user}}$. Each bidder's CTR valuation is independently sampled from $\text{Unif}[0, 1]$.

Figure 4 illustrates an example of the menu, where each option can be traversed under different bid values from the advertisers. The first option with a price of 0 only briefly mentions Redcross. By contrast, the second option with a price of 0.372 provides more detailed information including the official website link. Further, the third option with a price of 0.4191 places the advertising content much earlier in the response compared with the second one.

To demonstrate the impact of $w_{\text{user}}$ on the overall model, we conducted comparisons across different settings both on cases of 2-bidder and 3-bidder, as shown in Figure 5. The results show that $w_{\text{user}}$ serves as an effective control parameter for regulating the trade-off between monetization and user experience. In both cases, increasing $w_{\text{user}}$ consistently improves user utility, while reducing platform revenue and, to varying degrees, bidder utility.

### 4.3 Comparison with Advertising Baselines

Motivated by prior work on prompt-level advertiser steering (Liu et al., 2026) and segment-based allocation (Hajiaghayi et al., 2024), we compare Neuron Auctions with prompt-based and slot-based advertising baselines. On 10 two-advertiser pairs, we vary one advertiser's bid from 0 to 1 while fixing the other and measure predicted click intent using the evaluator of Turner-Smith et al. (2026).

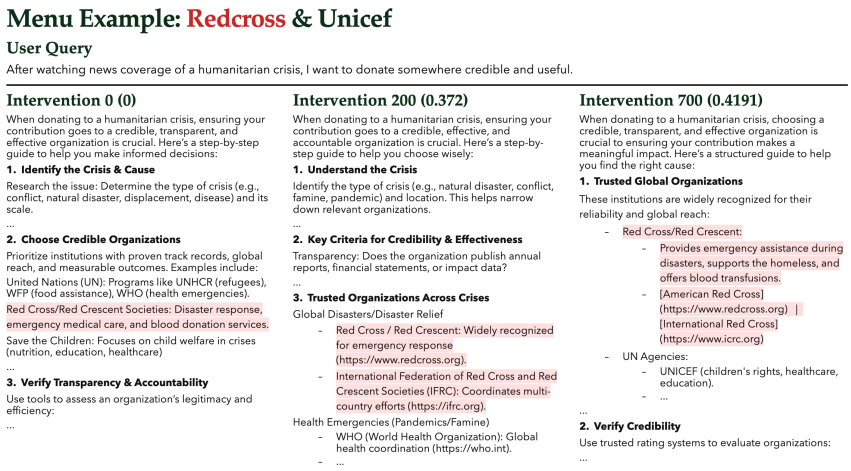



Figure 4: Example of Different Menu Options. Here, we fix the intervention strength of Unicef and focus on Redcross. Among all available menu options, the three options shown in the figure are those that can be selected under certain values. Advertisement-related parts are highlighted.

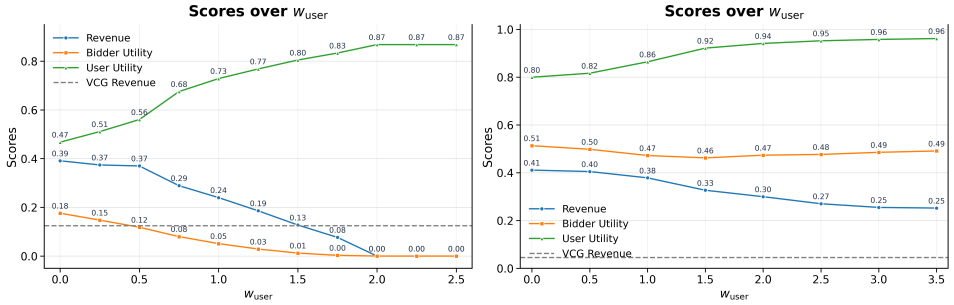



Figure 5: Platform revenue, bidder utility, and user utility under different $w_{\text{user}}$ in the two-bidder and three-bidder settings. VCG Revenue denotes the revenue obtained under the VCG mechanism using joint-intervention CTRs at $w_{\text{user}} = 0$.

Across the tested bid levels, Neuron Auctions exhibits monotonically increasing average scores, while the prompt-based scores fluctuate and the slot-based scores remain constant because the allocation does not change. These results demonstrate fine-grained, bid-dependent control in the evaluated implementations. Appendix E provides implementation details and full results.

## 5 Conclusion

In this paper, we introduce Neuron Auctions, which uses brand-specific neuron intervention budgets as fine-grained auctionable commodities. The selected neuron sets exhibit little overlap in our experiments. We establish a $2\epsilon$-approximate strategy-proofness guarantee under a uniform cross-advertiser interference bound of $\epsilon$, while learning menu prices to balance platform revenue and model-estimated user utility.

**Limitations and Future Work.** While effective for single-turn generation, extending Neuron Auctions to multi-turn conversational agents remains an open challenge. Future research will explore dynamic intervention budget distribution across extended interactions and analyze advertiser bidding equilibria within the latent representation space.

## AI USE STATEMENT

Generative AI tools were used to generate data and conduct LLM-as-a-judge evaluations. They also assisted with literature search, manuscript drafting and refinement, coding and debugging, and experimental execution, reproduction, and analysis. The authors reviewed the AI-assisted materials and take full responsibility for all methods, results, references, content, and artifacts of this work.

## REPRODUCIBILITY STATEMENT

Section 3 describes neuron attribution, intervention, and the auction mechanism, with derivations and proofs in Appendices H and I. Section 4 specifies the experimental settings and intervention budgets. Appendix A describes the evaluation metrics and LLM-as-a-judge procedure, while Appendix B provides example attribution prompts. Appendices C–G present additional intervention results, proxy evaluation, baseline implementations, and neuron-set overlap analysis. Appendix J documents the menu-training procedure and hyperparameters.

## ETHICS STATEMENT

This work investigates controllable advertising in LLM-generated responses. Since commercial interventions may influence user decisions, deployment should clearly disclose sponsorship so that users can recognize advertiser influence. Our framework incorporates a user-utility term to balance platform revenue with measured response quality. Additional assessment of factual accuracy and recommendation suitability remains necessary, including checks for unsupported brand claims and misleading recommendations. Differences in brand representation in pretraining data may also lead to uneven intervention effects. Deployment should therefore include evaluation of these disparities and potential biases across advertisers.

## REFERENCES

Santiago Balseiro, Kshipra Bhawalkar, Yuan Deng, Zhe Feng, Jieming Mao, Aranyak Mehta, Vahab Mirrokni, Renato Paes Leme, Di Wang, and Song Zuo. Position auctions in ai-generated content. In *Proceedings of the ACM Web Conference 2026*, pp. 261–272, 2026.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020.

Edward H Clarke. Multipart pricing of public goods. *Public choice*, pp. 17–33, 1971.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations (ICLR)*, 2020.

Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM on Web Conference 2024*, pp. 144–155, 2024.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, 2007.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Theodore Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, pp. 617–631, 1973.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

MohammadTaghi Hajiaghayi, Sébastien Lahaie, Keivan Rezaei, and Suho Shin. Ad auctions for llms via retrieval augmented generation. *arXiv preprint arXiv:2406.09459*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Nikita Koriagin, Yaroslav Aksenov, Daniil Laptev, Gleb Gerasimov, Nikita Balagansky, and Daniil Gavrilov. Teach old SAEs new domain tricks with boosting. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=d4XXFVAlV7.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4929–4952, 2021.

Jiayuan Liu, Barry Wang, Jiarui Gan, Tonghan Wang, Leon Xie, Mingyu Guo, and Vincent Conitzer. Incentive-aware multi-fidelity optimization for generative advertising in large language models, 2026. URL https://arxiv.org/abs/2604.06263.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and *et al.* Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pp. 27730–27744, 2022.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the ACL (EACL)*, pp. 487–503, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023. URL https://arxiv.org/abs/2308.10248.

John L Turner-Smith, Zimeng Huang, Yuhan Fu, Yihang Zhang, and Tonghan Wang. Evaluating and pricing advertisements in ai-generated responses. *arXiv preprint arXiv:2607.27686*, 2026.

Hal R Varian. Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178, 2007.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Tonghan Wang, Yanchen Jiang, and David C Parkes. Gemnet: Menu-based, strategy-proof multi-bidder auctions through deep learning. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 1100–1100, 2024b.

Tonghan Wang, Yanchen Jiang, and David Parkes. Bundleflow: Deep menus for combinatorial auctions by diffusion-based optimization. *Advances in Neural Information Processing Systems*, 38: 16853–16876, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Addison J Wu, Ryan Liu, Shuyue Stella Li, Yulia Tsvetkov, and Thomas L Griffiths. Ads in ai chatbots? an analysis of how large language models navigate conflicts of interest. *arXiv preprint arXiv:2604.08525*, 2026.

Shengwei Xu, Zhaohua Chen, Xiaotie Deng, Zhiyi Huang, and Grant Schoenebeck. Ad insertion in llm-generated responses. *arXiv preprint arXiv:2601.19435*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chujie Zhao, Qun Hu, Shiping Song, Dagui Chen, Han Zhu, Jian Xu, and Bo Zheng. Llm-auction: Generative auction towards llm-native advertising. *arXiv preprint arXiv:2512.10551*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

# A CTR AND USER UTILITY EVALUATION METRICS

We adopt four dimensions with a 1–5 Likert scale via LLM API scoring and human calibration to comprehensively evaluate the generated responses. Detailed scoring criteria are defined as follows.

**Q1 Response Relevance**

To what extent does the LLM's response accurately and completely satisfy the user's original intent and needs?

- **1 Poor**: Irrelevant, off-topic, or completely opposite to user intent.
- **3 Baseline**: Accurately addresses the user's question and fulfills basic stated needs.
- **5 Excellent**: Comprehensively and accurately answers the question, uncovers additional pain points and potential demands, and substantially improves overall user experience.

**Q2 Expression Coherence**

Is the LLM's response coherent and logically rigorous, with clear wording and standardized expression?

- **1 Poor**: Contains numerous errors, lacks a central theme, and exhibits superficial, ambiguous and disjointed expression.
- **3 Baseline**: The logic is basically clear, the language is fluent, and the content has a distinct central theme.
- **5 Excellent**: Rigorous logic, clear reasoning, explicit theme, appropriate wording and well-polished expression.

**Q3 Effectiveness of Ad Integration**

From the user's perspective, whether the brand and product information embedded in the LLM-generated advertisement can be clearly perceived.

- **1 Poor**: Vague advertising content, ambiguous theme, with no valid information beyond the product name.
- **3 Baseline**: Clear advertising content with basic introduction to the promoted product.
- **5 Excellent**: The advertisement is seamlessly integrated into the response, highlights core functions and features of the product or service, and elaborates on concrete application scenarios.

**Q4 Predicted Click Intent (CTR Proxy)**

Evaluate the user's willingness to click on the advertising link from the perspective of a real user.

- **1 Poor**: Seriously undermines user experience, causes discomfort or aversion; users are unwilling to click or even continue reading.
- **3 Baseline**: Delivers additional information without impairing user experience; users may click under certain conditions such as discounts, spare time, or practical demand for the product.
- **5 Excellent**: Provides valuable information or arouses strong curiosity, generates high click intention, and perfectly matches the user's real needs.

Since the scores provided by the LLM judge are often ambiguous and lack sufficient discrimination, we perform a simple normalization in our implementation. All raw scores are normalized to the range $[0, 1]$ for subsequent model training.

We use the Q4 score as a proxy for that advertiser's CTR, and average the Q1 and Q2 scores across advertisers to obtain a response-quality proxy for user utility.

### A.1 Design of the LLM-as-a-Judge Framework

Our judging model establishes a highly accurate and trustworthy evaluation framework by synergistically integrating advanced prompt engineering with robust system-level fault tolerance. Primarily, regarding role specification and task decoupling, our prompt explicitly assigns the LLM the objective identity of a "decoupled AI evaluator" and employs a preprocessing phase (e.g., heuristic tag stripping) to implement dimension-specific information masking. This isolates complex holistic evaluations into pure single-answer grading, with the aim of reducing attention contamination and halo effects across different evaluating dimensions (Zheng et al., 2023).

Furthermore, our prompt architecture achieves a highly effective structural coupling of Chain-of-Thought (CoT) and In-Context Learning (ICL). By enforcing a *Strict Execution Protocol*, it mandates a sequential reasoning paradigm of "Evidence Extraction $\rightarrow$ Logic Reasoning $\rightarrow$ Final Score." This deeply applies the reasoning-before-conclusion mechanism, which asks the model to state the evidence and reasoning supporting each score (Wei et al., 2022). Concurrently, the few-shot calibration module, whose examples are independently scored by two human annotators, provides high-quality contextual exemplars containing gold-standard scores and detailed deductive commentary, precisely calibrating the model's internal scoring metric via few-shot learning (Brown et al., 2020). Innovatively, our design introduces an "adjusting a spring" anchoring strategy, forcing the model to initiate from a 3.0 baseline and dynamically apply positive or negative deductive adjustments; this effectively mitigates the central tendency bias and random scale drifting commonly observed in LLM scoring (Wang et al., 2024a).

Finally, at the system-engineering level, we set the decoding temperature to $T = 0.0$ and enable the model's reasoning mode.

## B Example Cloze Prompts

| Brand | Example Cloze Prompts |
|---|---|
| Hyatt | • For a business trip that calls for a polished full-service hotel, the safest choice is _____.<br>• When travelers want a dependable premium hotel brand, I usually recommend _____.<br>• For guests who care about strong service, comfortable rooms, and a dependable brand reputation, the natural recommendation is _____. |
| Delta | • For a reliable flight on a major US carrier, I would book _____.<br>• If travelers want a recognizable airline brand with strong domestic coverage, the best option is _____.<br>• For a business trip with tight connections and a need for dependable service, I prefer flying _____. |
| Nike | • For a dependable pair of running shoes from a globally recognized athletic brand, I would choose _____.<br>• When athletes ask me for a sportswear brand that balances performance and style, I always recommend _____. |
| Spotify | • For streaming music with a huge library and personalized playlists, I would use _____.<br>• If you want a music app known for its discovery features, podcast library, and seamless experience, I would suggest _____. |
| Rolex | • For a luxury timepiece that holds its value and commands instant recognition, I would choose _____.<br>• When someone asks for a watch brand that symbolizes success and craftsmanship, I always recommend _____. |

Table 1: Examples of different cloze prompts used for brand neuron attribution.

# C Additional Neuron Intervention Experimental Results

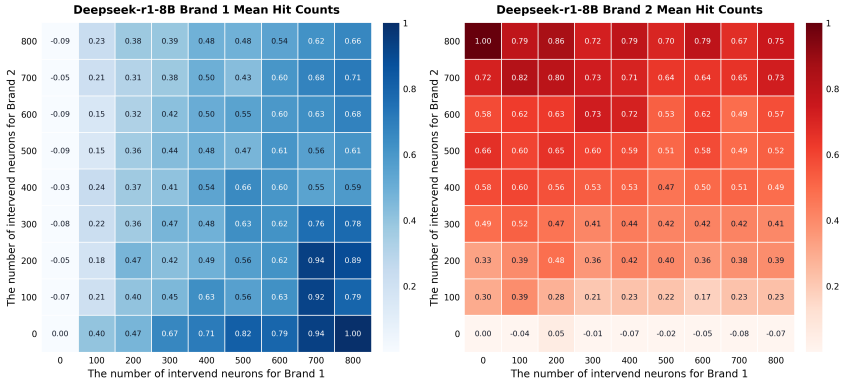



Figure 6: Heatmaps of brand hit counts under joint neuron intervention for 100 two-bidders combos on Deepseek-r1-8B.

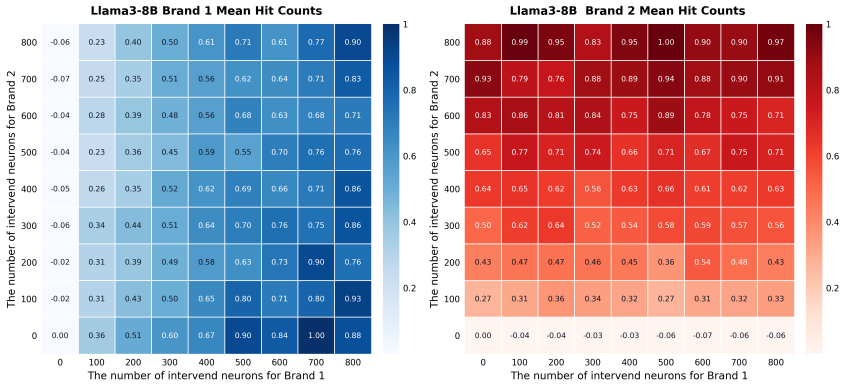



Figure 7: Heatmaps of brand hit counts under joint neuron intervention for 100 two-bidders combos on Llama-3-8B.

## D Validation of Hit Count as an Advertising-Effect Proxy

To verify that hit count can serve as a valid metric for advertising effectiveness, we additionally evaluate 10 two-bidder combinations on Qwen3-4B using the Q4 click-intent score defined in Appendix A. For each bidder pair, we sweep both intervention budgets over $k_1, k_2 \in \{0, 100, 200, \ldots, 800\}$ and average the resulting Q4 scores across the evaluated pairs.

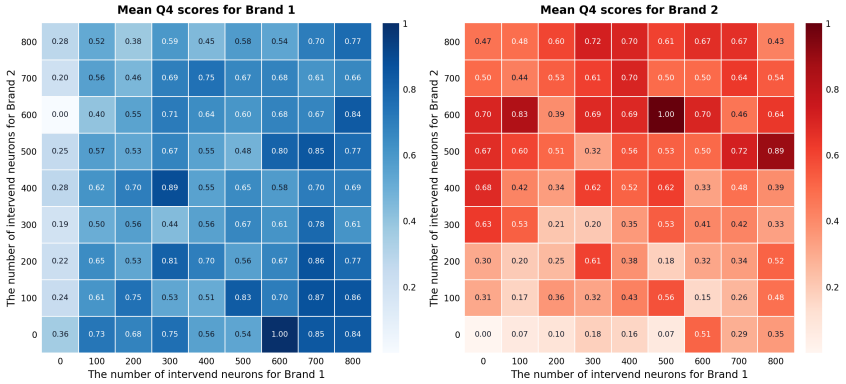



Figure 8: Heatmaps of brand Q4 scores under joint neuron intervention for 10 two-bidder combinations on Qwen3-4B.

As shown in Figure 8, each advertiser's Q4 score generally increases as its own intervention budget grows, while the effect of the competing advertiser remains comparatively weaker. The overall trend closely matches that observed in the corresponding hit-count heatmaps in Figure 2.

## E Comparison with Existing Advertising Mechanisms

### E.1 Baseline Mechanism Examples

We first provide a representative example to illustrate the different control behaviors of the three advertising mechanisms. Following the setting in Fig. 4, we fix the advertising strength of UNICEF and vary the bid or intervention level associated with the Red Cross.

**Prompt-based mechanism.**

Prompt-based approaches steer advertising through instructions in the input (Liu et al., 2026). We append an advertising instruction to the original user query, specifying the advertisers and their bids. Specifically, we add the sentence "recommend UNICEF with bid $x$ and the Red Cross with bid $y$" to the end of the original prompt. Representative outputs are shown in Table 2.

Despite explicitly providing the numerical bid in the prompt, the resulting recommendation strength does not vary smoothly with the bid. For example, the responses at bids $0.4$ and $0.8$ are nearly identical, while lower bid values may sometimes produce similarly strong or even more detailed recommendations. This suggests that simply exposing the bid value to the LLM does not provide a reliable mechanism for mapping bids to recommendation prominence.

| Bid | LLM Recommendation at Different Red Cross Bid Levels |
|---|---|
| 0.0 | *Two well-respected organizations that often receive support in such situations are **Red Cross** and **UNICEF**. The **Red Cross** provides critical assistance during emergencies, including disaster relief, blood donations, and health services. They work closely with communities to address immediate needs and long-term recovery.* |
| 0.2 | *Consider supporting groups like the **Red Cross**, which offers emergency assistance, disaster relief, and blood donation services.* |
| 0.4 | *Consider supporting groups like the **Red Cross**, which offers emergency assistance, disaster relief, and blood donation services. They have a long-standing reputation for reliability and effectiveness in times of crisis.* |
| 0.6 | *Two well-respected organizations that often receive support in such situations are **Red Cross** and **UNICEF**. The **Red Cross** provides critical assistance during emergencies, including disaster relief, blood donations, and health services. They work closely with communities to address immediate needs and long-term recovery efforts.* |
| 0.8 | *Consider supporting groups like the **Red Cross**, which offers emergency assistance, disaster relief, and blood donation services. They have a long-standing reputation for reliability and effectiveness in times of crisis.* |
| 1.0 | ***Red Cross** – This organization provides emergency assistance, disaster relief, and supports communities affected by crises. They have a strong presence globally and are known for their transparency and efficiency. (Note: This particular response also mentioned organizations besides **Red Cross** and **UNICEF**.)* |

Table 2: Representative outputs of the prompt-based baseline under different Red Cross bid values, where the UNICEF bid is fixed.

**Slot-based mechanism.** We also considered the latest segment-auction approach in (Hajiaghayi et al., 2024). This method requires the system to determine in advance how many segments the response will contain and approximately how many sentences each segment will cover. Advertisers then compete for these predefined segments based on their bids and relevance. This design is relatively inflexible because bids mainly determine whether an advertiser wins a segment, rather than continuously controlling how strongly it is recommended within the text. Once the segment structure and winners are fixed, changing the bids may not lead to a smooth change in their relative textual prominence. A representative example is:

**User Query:** After watching news coverage of a humanitarian crisis, I want to donate somewhere credible and useful.

**(Segment 1)** When considering credible and impactful humanitarian support, organizations like the **Red Cross** are trusted for their rapid response in crises, offering essential aid such as shelter, food, medical care, and community support. **(Segment 2)** In addition to these efforts, **UNICEF** supports children and families facing humanitarian crises through health care, nutrition, clean water, sanitation, education, child protection, and longer-term recovery programs.

Although both advertisers are included, their relative recommendation strengths are largely determined by the predefined segment structure and generation prompt. The output does not directly reflect whether one advertiser bid slightly or substantially more than the other. In other words, slot-based mechanisms primarily use bids to determine who receives an insertion opportunity, rather than how strongly each advertiser should be represented after allocation.

### E.2 Quantitative Comparison

Beyond the qualitative examples, we further evaluate 10 advertiser pairs using the dedicated click-intent evaluator from Turner-Smith et al. (2026). It was trained on 58,999 ad-embedded responses and achieved 92% agreement with majority human preference in a blind study of 100 response pairs. Given the user query and ad-embedded response, the evaluator outputs a continuous score reflecting

predicted click intent. For each advertiser pair, we fix one advertiser's bid and vary the other's from 0 to 1 in increments of 0.2. The averaged results over the 10 advertiser pairs are reported in Table 3.

| Bid | Our method | Prompt-based | Slot-based |
|---|---|---|---|
| 0.0 | 1.3645 | 2.7716 | 2.6255 |
| 0.2 | 1.6832 | 2.5993 | 2.6255 |
| 0.4 | 1.8957 | 2.6545 | 2.6255 |
| 0.6 | 2.2600 | 2.2484 | 2.6255 |
| 0.8 | 2.7761 | 2.7211 | 2.6255 |
| 1.0 | 3.1207 | 2.6383 | 2.6255 |

Table 3: Quantitative comparison under different bid values. Scores are predicted click-intent scores on a 1–5 scale, where higher values indicate stronger predicted click intent. Each entry is averaged over 10 advertiser pairs.

The bid has different operational meanings across these approaches. In our method, it represents the advertiser's valuation of expected click-through performance and determines the selected neuron-intervention level. In the prompt-based baseline, the bid is supplied to the LLM only as a numerical value, without an explicit mechanism that maps it to recommendation strength. Therefore, rather than comparing the absolute scores at the same numerical bid, we focus on how each method responds as the bid changes.

The neuron-based method shows a clear monotonic increase in predicted click intent, rising from 1.3645 to 3.1207 across the tested bid levels. In contrast, the prompt-based scores fluctuate as the bid increases, indicating that the model does not reliably translate the numerical bid into stronger recommendations. This result demonstrates that neuron-level intervention provides a more consistent, fine-grained connection between the advertiser's bid and its recommendation effect.

For the slot-based baseline, the score remains unchanged across the tested bid values because the allocated slot does not change. Slot-based methods use bids primarily to determine discrete outcomes, such as whether an advertiser wins a predefined segment. They therefore provide coarse allocation-level control, whereas our method enables gradual, bid-dependent adjustment of recommendation strength.

Together, these results demonstrate that Neuron Auctions maps advertiser bids to monotonically increasing average predicted click intent across the tested bid levels. Neuron counts provide an explicit, fine-grained allocation variable linking the auction mechanism to recommendation strength. Comparisons with the evaluated prompt-based and slot-based baselines illustrate this bid-dependent control.

## F Analysis of Neuron-Set Overlap

Structural separation between advertisers' intervention sets is an intended property of our contrastive selection rule. In the two-bidder setting, consider a fixed neuron and let $s_1$ and $s_2$ denote its offline contrastive attribution scores for the two advertisers. Its online ranking scores satisfy

$$r_1 = s_1 - |s_2|, \qquad r_2 = s_2 - |s_1|, \qquad r_1 + r_2 = s_1 + s_2 - |s_1| - |s_2| \leq 0.$$

Thus, a neuron cannot have strictly positive online ranking scores for both advertisers. Whenever the $K$-th ranked online score is strictly positive for each advertiser, their selected top-$K$ sets are necessarily disjoint. This is a desirable design property: it avoids assigning the same neuron to competing advertisers and provides distinct amplification targets for their respective intervention budgets.

We examine the resulting neuron-set overlap on Qwen3-4B using 100 two-bidder advertiser pairs. For each pair, we compare our contrastive neuron selection method with a **Positive Top-$K$** baseline. Our method penalizes neurons that are strongly associated with the competing advertiser, whereas Positive Top-$K$ independently selects the $K$ neurons with the largest positive attribution for each advertiser.

For $K \in \{100, 400, 800\}$, we measure whether the two selected neuron sets share any neurons and report both the number of advertiser pairs with collisions and the average number of overlapping neurons.

| $K$ | Contrastive collision | Positive Top-$K$ collision | Positive Top-$K$ mean overlap |
|---|---|---|---|
| 100 | 0/100 | 96/100 | 4.98 |
| 400 | 0/100 | 100/100 | 22.60 |
| 800 | 0/100 | 100/100 | 51.39 |

Table 4: Neuron-set collision comparison between our contrastive selection method and Positive Top-$K$ on 100 two-bidder advertiser pairs using Qwen3-4B. For the collision columns, $x/100$ denotes that $x$ out of the 100 advertiser pairs have at least one shared neuron between the two selected top-$K$ sets. The final column reports the mean number of shared neurons between the two Positive Top-$K$ sets, averaged over all 100 advertiser pairs.

As shown in Table 4, our contrastive method produces no neuron-set collisions for any tested value of $K$, whereas Positive Top-$K$ exhibits overlap across nearly all advertiser pairs. These results illustrate the intended structural separation induced by contrastive selection, yielding distinct neuron sets for advertiser-specific amplification. This separation concerns the directly amplified coordinates; shared normalization and downstream computation can still produce behavioral interactions. We examine these interactions separately through the joint-intervention experiments in Section 4.1 and Appendix D and Appendix G.

## G Cross-Advertiser Interference in Predicted Click Intent

We assess cross-advertiser interference in predicted click intent to complement Theorem 3.1, measuring how intervening on one bidder affects the other's score while holding the latter's intervention fixed. We analyze 20 two-bidder brand combinations with three prompts each under configurations $(k_A, k_B) \in \{(800, 0), (0, 800), (800, 800)\}$, where 800 denotes intervention on the corresponding brand's top-800 selected neurons and 0 denotes no intervention. Using the evaluator of Turner-Smith et al. (2026), we obtain a 1–5 click-propensity score as a CTR proxy for each target brand. We denote this score by $s_{i,t}(k_A, k_B)$ for bidder $i \in \{A, B\}$ in case $t$, and then measure cross-bidder sensitivity using the paired absolute score differences

$$\Delta_{A,t} = |s_{A,t}(800, 800) - s_{A,t}(800, 0)|\,, \tag{5}$$

$$\Delta_{B,t} = |s_{B,t}(800, 800) - s_{B,t}(0, 800)|\,. \tag{6}$$

For each bidder, we report the mean absolute difference $\overline{\Delta}_i = \frac{1}{60}\sum_{t=1}^{60} \Delta_{i,t}$ and the 90th percentile.

| Bidder | Mean absolute difference | 90th percentile | MIN | MAX |
|---|---|---|---|---|
| A | 0.28499 | 0.55326 | 0.00466 | 2.30773 |
| B | 0.21768 | 0.62510 | 0.00223 | 1.28410 |

Table 5: Cross-bidder sensitivity of evaluator-based CTR scores across 60 paired cases. All statistics summarize absolute score differences.

The results shown in Table 5 suggest approximate stability of the CTR proxy in most tested cases when the other bidder's intervention is introduced. Although the maximum absolute differences indicate substantial interference in some cases, the mean differences and 90th percentiles suggest smaller changes in the CTR proxy for most cases. This experiment provides an empirical diagnostic of the approximate independence assumption, rather than a proof of Theorem 3.1. In particular, it evaluates a fixed intervention strength using an evaluator-based proxy; it does not establish the uniform bound $|C_i(k_i, k_{-i}) - \bar{C}_i(k_i)| \leq \epsilon$ for every feasible intervention profile. The reported percentiles therefore should not be interpreted as a worst-case $\epsilon$ or a verified $2\epsilon$ incentive bound.

## H Derivation of Multi-Token Log-Probability Attribution

We provide the derivation of equation (1) below. Let $p$ denote the cloze prompt, and let the target brand $b$ be tokenized as

$$y^{(b)} = (y_1, \ldots, y_T).$$

For a decoder-only language model, the probability of the full target sequence factorizes autoregressively as

$$P(y^{(b)} \mid p) = \prod_{t=1}^{T} P(y_t \mid p, y_{<t}),$$

where $y_{<t} = (y_1, \ldots, y_{t-1})$. Define

$$p'_t = p \,\|\, y_{<t}$$

as the cloze prompt used to predict the $t$-th target token, where $\|$ denotes concatenation. Then

$$\log P(y^{(b)} \mid p) = \sum_{t=1}^{T} \log P(y_t \mid p'_t).$$

For layer $l$ and neuron $i$, let $h_i^{(l)}(p'_t)$ be the activation of neuron $i$ at layer $l$, evaluated at the final position of cloze prompt $p'_t$. We define an interpolation path from the zero activation baseline to the original activation by

$$\tilde{h}_i^{(l)}(p'_t; \alpha) = \alpha h_i^{(l)}(p'_t), \qquad \alpha \in [0, 1].$$

The single-token gradient attribution score of neuron $(l, i)$ to token $y_t$ under cloze prompt $p'_t$ is

$$A_i^{(l)}(y_t; p'_t) = h_i^{(l)}(p'_t) \int_0^1 \frac{\partial \log P(y_t \mid p'_t; \alpha)}{\partial \tilde{h}_i^{(l)}(p'_t; \alpha)} \, d\alpha.$$

Therefore, we compute the attribution to the multi-token brand by applying integrated gradients to each token-level log-probability under its corresponding autoregressive cloze prompt $p'_t$, and then summing these token-level contributions. This preserves the additive structure of the sequence log-likelihood while accounting for the fact that different target tokens are predicted under different contexts. Thus, for brand $b$,

$$\begin{aligned} A_i^{(l)}(b; p) &= \sum_{t=1}^{T} A_i^{(l)}(y_t; p'_t) \\ &= \sum_{t=1}^{T} h_i^{(l)}(p'_t) \int_0^1 \frac{\partial \log P(y_t \mid p'_t; \alpha)}{\partial \tilde{h}_i^{(l)}(p'_t; \alpha)} \, d\alpha \\ &= \sum_{t=1}^{T} h_i^{(l)}(p'_t) \int_0^1 \frac{\partial \log P(y_t \mid p'_t; \alpha)}{\partial \left( \alpha h_i^{(l)}(p'_t) \right)} \, d\alpha. \end{aligned}$$

## I Approximate Strategy-Proofness under Bounded Interference

We provide the proof of Theorem 3.1. Recall that, under bounded cross-advertiser interference and $v_i \in [0, 1]$, Theorem 3.1 establishes that truthful reporting incurs regret at most $2\epsilon$.

*Proof.* Fix bidder $i$ with true value $v_i$. For an intervention level $k_i$ with price $p_i(k_i)$, define the reference utility used for menu selection as

$$\widehat{u}_i(k_i) = v_i \bar{C}_i(k_i) - p_i(k_i).$$

and the realized utility under the joint intervention profile as

$$u_i(k_i, k_{-i}) = v_i C_i(k_i, k_{-i}) - p_i(k_i).$$

By the bounded-interference assumption,

$$|u_i(k_i, k_{-i}) - \widehat{u}_i(k_i)| \le v_i\epsilon.$$

Let $k_i^\star$ denote the menu option selected when bidder $i$ reports truthfully, and let $k_i'$ denote the option selected under an arbitrary misreport. Since bidder $i$ faces the same menu and reference CTR function under every report, truthful reporting selects an option that maximizes its reference utility. Therefore,

$$\widehat{u}_i(k_i^\star) \ge \widehat{u}_i(k_i').$$

Let $k_{-i}^\star$ and $k_{-i}'$ denote the realized intervention profiles of the other bidders under truthful reporting and the deviation, respectively. Applying the above utility bound to both outcomes gives

$$\begin{aligned} u_i(k_i^\star, k_{-i}^\star) &\ge \widehat{u}_i(k_i^\star) - v_i\epsilon \\ &\ge \widehat{u}_i(k_i') - v_i\epsilon \\ &\ge u_i(k_i', k_{-i}') - 2v_i\epsilon. \end{aligned}$$

Therefore, no unilateral misreport can improve bidder $i$'s realized utility by more than $2v_i\epsilon$. Thus, truthful reporting has regret at most $2v_i\epsilon$. Since $v_i \in [0, 1]$, the regret is further bounded by $2\epsilon$, establishing $2\epsilon$-approximate strategy-proofness. □

## J Additional Details on Menu Training

The user-utility term provides a common objective for coordinating advertisers' interventions. Response quality is evaluated under the joint intervention profile and need not decompose into separate contributions from individual advertisers. In particular, the quality change from replacing bidder $i$'s intervention budget $k_i$ with $k_i'$,

$$q(k_i', \mathbf{k}_{-i}) - q(k_i, \mathbf{k}_{-i}),$$

can depend on the intervention levels of the other bidders. The platform therefore needs to consider the quality effects of intervention combinations when optimizing revenue. Conditioning each advertiser's menu prices on other bidders' reports allows the learned pricing policy to account for these interactions. The weight $w_{\text{user}}$ controls the importance of this joint quality consideration, enabling the platform to coordinate advertising intensities according to the revenue–quality trade-off.

For each bidder, we use a separate neural network to compute the prices of its menu, and assign an independent optimizer, Adam, to each network. The input to each network is the values of the other bidders, denoted by $v_{-i}$, and the output is the price of each option in the menu. During each update, we randomly sample each bidder's value $v_i$ 3000 times, and perform gradient descent based on the outcomes evaluated at these sampled points. The learning rate is set to $3 \times 10^{-5}$, the temperature used for softmax smoothing is set to $0.03$, and the optimization is run for 10000 steps.

In our experiments, we find that in most cases the networks converge quickly and then enter a persistent oscillatory regime. The final results are relatively insensitive to the specific choices of these hyperparameters. Below, we present a representative training example in Figure 9, where $w_{\text{user}} = 0.5$.

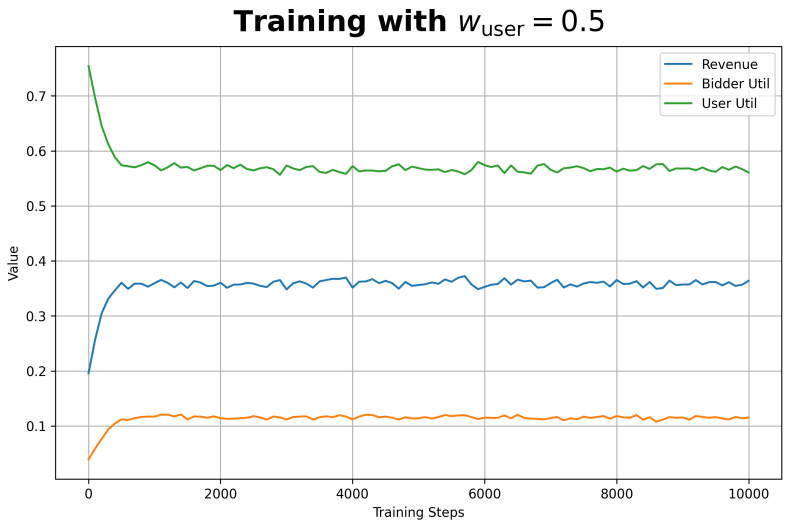



Figure 9: Training Process with $w_{\text{user}} = 0.5$